GATE — General Architecture for Text Engineering

THE UNIVERSITY OF SHEFFIELD
Department of
computer science

DAML+OIL Export

Last updated: Friday, 7 February 2003

sheffieldnlp

natural language processing group

1. Introduction

The DAML+OIL Export is a GATE PR that allows the named entities found in documents to be exported as instances of a specified ontology in DAML+OIL format. At present only DAML+OIL (http://www.daml.org) is supported, but migrating the code to OWL (http://www.w3.org/TR/owl-ref/) is trivial.

The DAML+OIL Export can work in two modes. In the first mode (using the normal gazetteer), all you need is to have an ontology containing concepts such as Person, Location, Organization, etc. corresponding to the named entity types recognized in GATE. When you have a corpus processed with ANNIE (so that certain named entities in the corpus are recognized) then you can create a DAML+OIL Export processing resource (specifying as initialisation parameter the ontology to be used as reference) and when the DAML+OIL resource processes the (already annotated) corpus, for each named entity found that is of some type (such as Location), if a corresponding concept with the same name as the named entity type (such as Location) exists in the ontology then a new DAML instance will be generated in the export file (f.e. <gate:Location rdf:about="…"/>).

The second way to use the DAML+OIL export is when you have used the OntoText OntoGazetteer[1] (instead of the default ANNIE gazetteer) to annotate the corpus. The OntoGazetter will works in a way similar to the default gazetteer but it will generate more meaningful annotations according to some ontology, i.e. instead of having a Location annotation the OntoGazetteer may generate more specific annotations such as City, River, Mountain, etc. When the DAML+OIL Export processes a corpus that was annotated with the help of the OntoGazetteer, the exported instances will also be from the more specific types (such as City, Mountain, etc).

2. Using the DAML+OIL Export

1. Exporting a corpus annotated with the default gazetteer

The following steps should be performed:

1. Process the corpus with ANNIE
2. Create an ontology that will be used as reference from the DAML+OIL Export
3. Load a DAML+OIL PR in the GATE IDE, specifying the relevant ontology
4. Process the corpus with the DAML+OIL PR

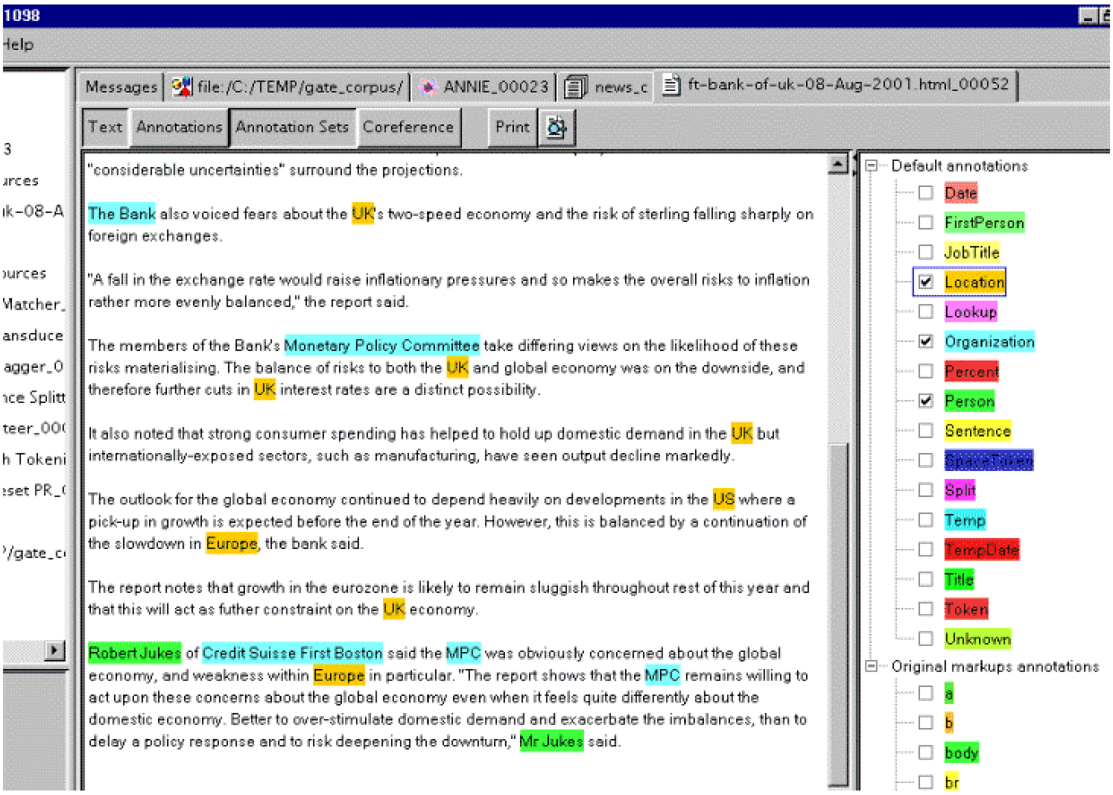
As a result, the specified output directory will contain DAML files representing instance data found in each document of the processed corpus

The following section presents a detailed explanation of the steps that should be followed.

1.1. Process the corpus with ANNIE

No special actions should be performed at this step. The corpus is annotated with ANNIE in the ordinary way. A sample processed file looks like:

[1] Details about the OntoGazetteer are available in the GATE Users Guide, section 4.1



"considerable uncertainties" surround the projections.

The Bank also voiced fears about the UK's two-speed economy and the risk of sterling falling sharply on foreign exchanges.

"A fall in the exchange rate would raise inflationary pressures and so makes the overall risks to inflation rather more evenly balanced," the report said.

The members of the Bank's Monetary Policy Committee take differing views on the likelihood of these risks materialising. The balance of risks to both the UK and global economy was on the downside, and therefore further cuts in UK interest rates are a distinct possibility.

It also noted that strong consumer spending has helped to hold up domestic demand in the UK but internationally-exposed sectors, such as manufacturing, have seen output decline markedly.

The outlook for the global economy continued to depend heavily on developments in the US where a pick-up in growth is expected before the end of the year. However, this is balanced by a continuation of the slowdown in Europe, the bank said.

The report notes that growth in the eurozone is likely to remain sluggish throughout rest of this year and that this will act as futher constraint on the UK economy.

Robert Jukes of Credit Suisse First Boston said the MPC was obviously concerned about the global economy, and weakness within Europe in particular. "The report shows that the MPC remains willing to act upon these concerns about the global economy even when it feels quite differently about the domestic economy. Better to over-stimulate domestic demand and exacerbate the imbalances, than to delay a policy response and to risk deepening the downturn," Mr Jukes said.

Default annotations
- ☐ Date
- ☐ FirstPerson
- ☐ JobTitle
- ☑ Location
- ☐ Lookup
- ☑ Organization
- ☐ Percent
- ☑ Person
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ Temp
- ☐ TempDate
- ☐ Title
- ☐ Token
- ☐ Unknown

Original markups annotations
- ☐ a
- ☐ b
- ☐ body
- ☐ br

1.2. Create an ontology

Create a DAML+OIL ontology that contains concepts such as Location, Person, Organization, etc. with names corresponding to the Named Entity types in GATE. This ontology will be used as reference from the DAML+OIL Export, i.e. for each annotation found in the processed corpus, the exporter will lookup the ontology for a concept with the same name and if such exists then an instance of this type will be generated in the output DAML file (containing instance data)

Example ontology is shown below. Note that the ontology contain much more concepts that necessary (such as City, Mountain, etc) because the corpus is annotated with the default gazetteer that will never generate such annotations.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
        xmlns="http://pillango.sirma.bg/gate#">
    <daml:Ontology rdf:about="">
        <daml:versionInfo>1.0</daml:versionInfo>
        <daml:imports rdf:resource="http://www.daml.org/2001/03/daml+oil" />
    </daml:Ontology>

    <rdfs:Class rdf:ID="Businessman">
        <rdfs:subClassOf rdf:resource="#Person" />
    </rdfs:Class>

    <rdfs:Class rdf:ID="Person" />

    <rdfs:Class rdf:ID="Organization" />
```

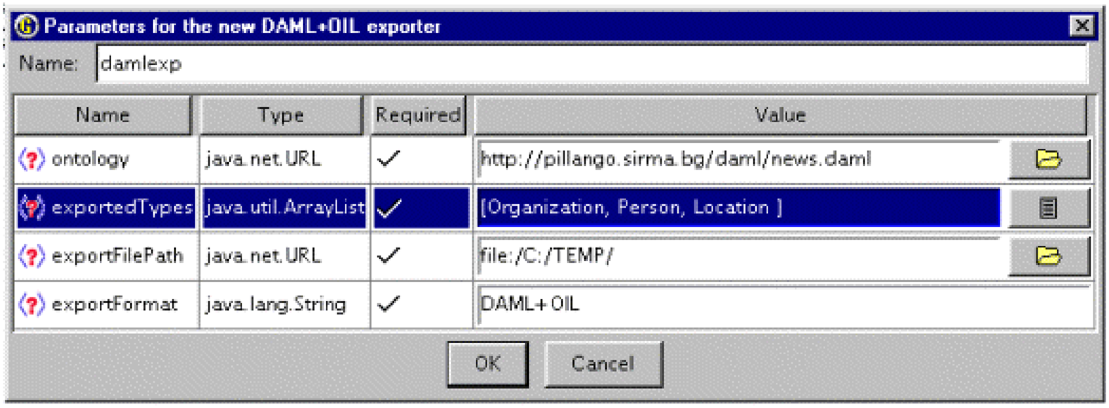
```
  <rdfs:Class rdf:ID="City">
      <rdfs:subClassOf rdf:resource="#Location" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Location" />

  <rdfs:Class rdf:ID="Company">
      <rdfs:subClassOf rdf:resource="#Organization" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Country">
      <rdfs:subClassOf rdf:resource="#Location" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Date" />

  <rdfs:Class rdf:ID="Department">
      <rdfs:subClassOf rdf:resource="#Government" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Government">
      <rdfs:subClassOf rdf:resource="#Organization" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="MediaPerson">
      <rdfs:subClassOf rdf:resource="#Person" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Ministry">
      <rdfs:subClassOf rdf:resource="#Government" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="MoneyAmount" />

  <rdfs:Class rdf:ID="Politician">
      <rdfs:subClassOf rdf:resource="#Person" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Province">
      <rdfs:subClassOf rdf:resource="#Location" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Region">
      <rdfs:subClassOf rdf:resource="#Location" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Sportsman">
      <rdfs:subClassOf rdf:resource="#Person" />
  </rdfs:Class>

  <rdfs:Class rdf:ID="Mountain">
      <rdfs:subClassOf rdf:resource="#Region" />
  </rdfs:Class>

</rdf:RDF>
```

1.3. Load the DAML+OIL processing resource in GATE

DAML+OIL Export is available in the default GATE distribution so you don't need to modify *creole.xml* in order to use it. When loading the resource one should specify:

- The ontology to be used

- The named entity types that will be exported (in this case only Organization, Person and Location annotations will be considered for export)
- Specify the output directory
- Choose the output format. At present only DAML+OIL is supported but other ontology languages such as OWL may be supported in the future.



Parameters for the new DAML+OIL exporter

Name: damlexp

Name	Type	Required	Value
ontology	java.net.URL	✓	http://pillango.sirma.bg/daml/news.daml
exportedTypes	java.util.ArrayList	✓	[Organization, Person, Location]
exportFilePath	java.net.URL	✓	file:/C:/TEMP/
exportFormat	java.lang.String	✓	DAML+OIL

In out example the *news.daml* ontology is the same ontology from step 1.2

1.4. Process the corpus with the DAML+OIL PR

Create a Corpus Pipeline containing inly the DAML+OIL Export processing resource and run it over the annotated corpus from step 1.1

Example output (for the file from 1.1) looks like:

```
 <?xml version="1.0" ?>
<rdf:RDF xmlns:gate="http://pillango.sirma.bg/daml/news.daml#"
      xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <daml:Ontology rdf:about="" daml:versionInfo="1.0">
      <daml:comment>autogenerated from GATE RDFFormatExporter</daml:comment>
   </daml:Ontology>
   <daml:Property rdf:about="http://www.daml.org/2001/03/daml+oil#comment" />
   <daml:Property rdf:about="http://www.daml.org/2001/03/daml+oil#versionInfo" />
   <daml:Property rdf:about="http://www.daml.org/2001/03/daml+oil#sameIndividualAs" />

   <gate:Location rdf:about="US" />

   <gate:Location rdf:about="UK" />

   <gate:Location rdf:about="Europe" />

   <gate:Organization rdf:about="Credit Suisse First Boston" />

   <gate:Organization rdf:about="The Financial Times" />

   <gate:Organization rdf:about="Bank of England">
      <daml:sameIndividualAs rdf:resource="The Bank"
           rdf:type="http://pillango.sirma.bg/daml/news.daml#Organization" />
   </gate:Organization>

   <gate:Organization rdf:about="MPC">
      <daml:sameIndividualAs rdf:resource="Monetary Policy Committee"
           rdf:type="http://pillango.sirma.bg/daml/news.daml#Organization" />
   </gate:Organization>
```

```
  <gate:Person rdf:about="Chris Flood" />

  <gate:Person rdf:about="Andrew Child" />

  <gate:Person rdf:about="Mr Jukes">
      <daml:sameIndividualAs rdf:resource="Robert Jukes"
           rdf:type="http://pillango.sirma.bg/daml/news.daml#Person" />
  </gate:Person>

</rdf:RDF>
```

Important facts:

- The namespace for the exported instance is *gate*
- Co-referring entities (identified by the Orthomatcher) will be linked with the help of the *daml:dameIndividualAs* construct
- Pronominal coreferents will not be exported, i.e. pronouns refering to entities in the text (he, she, etc) won't appear in the output

NOTE: the URL pillango.sirma.bg/daml/news.daml is used only for clarification. It should not be referenced in your applications

2. Exporting a corpus annotated with the OntoGazetteer

The steps for using the DAML+OIL Export with the OntoGazetteer are the same. The only difference is that in step 1.1 instead of using the ANNIE gazetteer, the corpus should be annotated with the OntoGazetteer and because it will generate more meaningful annotations such as City, River, etc. according to the specified ontology, the Export will also generate more specific instances. In our example, instead of having

```
<gate:Location rdf:about="US" />

<gate:Location rdf:about="UK" />
```

…instances, it is expected that

```
<gate:Country rdf:about="US" />

<gate:Country rdf:about="UK" />
```

…instances are exported, since the OntoGazetteer is expected to generate 'Country' annotations for "US" and "UK", instead of "Location" annotations

3. Using the DAML+OIL Export from Java

[TBD]